Evaluating Local LLMs on Japanese National University Entrance Examination Dataset in Comparison with Student Performance

Kyosuke Takami
Osaka Kyoiku University
takami-k75@cc.osaka-kyoiku.ac.jp

Satoshi Sekine
NII LLMC
sekine@nii.ac.jp

Yusuke Miyao
The University of Tokyo
NII LLMC
yusuke@is.s.u-tokyo.ac.jp

ABSTRACT

The evaluation of Large Language Models (LLMs) in educational contexts has largely focused on English-language datasets, limiting their applicability to non-English assessments. This study introduces a structured transformation of the Question Classification Annotation Specification dataset—a localized Japanese dataset derived from real-world university entrance exam (Center Test) questions—into a benchmark for LLM evaluation. A key feature of this dataset is the inclusion of average student scores, enabling direct comparison between LLMs and human test-takers. In this study, as preliminary investigation, we evaluated relatively small-scale LLMs (~13 billion parameters), particularly those developed with local linguistic resources such as Japanese (e.g., llm-jp, gemma3, DeepSeek R1 distill), to assess their feasibility for deploy-ment in local computing environments. By comparing these smaller-scale LLMs directly with human examinees, we found that, although most models did not reach average human performance, the exact match rate obtained by a certain model was considered comparable to human-level performance. This result implies that locally deployable 13B-scale LLMs have the potential to simulate human performance effectively.

Keywords

Comparing Human and Artificial Intelligence, Evaluation local LLM, Japanese National University Entrance Examination dataset

1. INTRODUCTION

With recent advancements in artificial intelligence, large language models have demonstrated considerable potential in educational contexts. However, the effectiveness of local LLMs—customized language models tailored to specific datasets—remains underexplored, particularly in evaluating student performance. This research addresses this gap by examining the applicability of local LLMs using the National University Entrance Examination dataset.

2. RELATED WORK

2.1 Applications of LLMs in Education

Large Language Models (LLMs) are being explored for various applications in educational settings. For example, models like ChatGPT demonstrate scores around the passing threshold for various standardized tests, including professional exams, though they struggle with complex reasoning tasks [14]. These capabilities have led to proposals for leveraging LLMs in student self-study, including generating test questions and administering mock exams. Some studies have reported attempts to simulate incorrect answers based on learners' response patterns to diagnose misconceptions and weaknesses [12]. Additionally, integrating LLMs with traditional cognitive diagnostic models has been proposed to enhance the accuracy of student proficiency assessment [2].

2.2 Japanese LLM Evaluation Benchmarks

Historically, LLM evaluation has centered on English, but Japanese and other non-English benchmarks have been developed. A notable example is JGLUE, which evaluates Japanese language comprehension and is constructed independently of English datasets [7]. Additionally, Japanese researchers have introduced the Nejumi leaderboard [10] integrating llm-jp-eval [4], which aggregates evaluations from 12 Japanese-language datasets and multi-turn Q&A datasets (Japanese MT Bench) to assess LLMs' comprehension and generative capabilities. Furthermore, domain-specific benchmarks, including those for Japanese medical licensing exams [6] and biomedical domains [5], have been established to assess the proficiency of Japanese LLMs.

2.3 Benefit of local LLM in Education

Cloud-based AI services require sending sensitive educational data (e.g. student responses, grades) to external servers, which can conflict with privacy regulations and trust. These concerns highlight the significance of local LLM deployments, where models run on-premises (e.g. school servers or teacher's devices) rather than on a remote cloud. Recent advances in model compression and hardware efficiency are making local deployment increasingly feasible [13]. These technologies will enable that "leaner" LLMs can be stored and run on ordinary laptops or phones with minimal performance loss, improving privacy, reducing latency, and lowering costs. Therefore, this study focuses on LLMs with approximately 13 billion parameters that can be executed within local computing environments. Additionally, we focused on LLMs known for their strength in local languages to investigate how well these models perform on tests conducted in the local language, and how their performance compares with that of human examinees.

Kyosuke Takami, Satoshi Sekine, and Yusuke Miyao. Evaluating Local LLMs on Japanese National University Entrance Examination Dataset in Comparison with Student Performance. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 582–585. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870238

3. DATASET

3.1 The National Center Test for University Admissions

The National Center Test for University Admissions (NCTUA), known in Japanese as "Daigaku Nyūshi Sentā Shiken," was a standardized examination administered annually in Japan to assess the academic abilities of prospective university students. This test played a pivotal role in the university admissions process, serving as a common benchmark for evaluating applicants' foundational knowledge across various subjects i.e. Physics, Biology, Chemistry, Mathematics, English, Japanese, History and so on. These exams consist of multiple-choice questions answered on mark sheets, with each subject test lasting approximately one to one and a half hours. It was administered from 1990 until its discontinuation in the 2020 academic year. From the 2021 academic year onward, a similar test has continued to be administered under the name "Common Test for University Admissions." called in Japanese as “Daigaku Nyūgaku Kyōtsū Tesuto”. The number of examinees and the average scores for each subject in this examination are published annually, making it possible to directly compare the performance of LLMs with that of high school students.

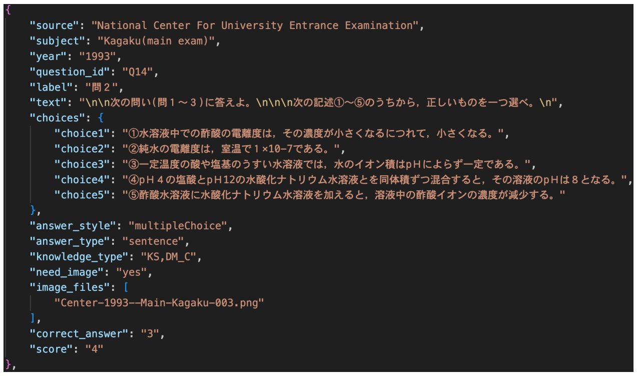
3.2 Previous XML Dataset

In this study, we utilized XML data of the National Center Test for University Admissions [11], prepared as part of the "Todai Robot Project", which explores whether artificial intelligence can pass the entrance examination for the University of Tokyo—the institution widely regarded as the most difficult to gain admission to in Japan. This XML dataset is carefully and faithfully structured, accurately capturing the details of exam questions—including specific elements such as underlined text and diagrams. Additionally, each question is annotated to indicate whether answering it requires external knowledge or whether it involves referencing and comprehending materials (texts) within the question itself. Utilizing this richly annotated data, we constructed a structured dataset optimized for evaluating language models. Specifically, we transformed the extensive metadata annotations to clearly identify knowledge requirements, enabling precise assessments of LLM performance—whether through external knowledge or question-internal comprehension. When preparing the dataset, the underlined sections were enclosed in brackets "【】". For cases where a main passage is followed by related sub-questions, we conducted specialized data preparation tailored for LLM evaluation by inserting the main text at the beginning of each corresponding question. And if an image is included in main text, we also added the image to each corresponding question. Through these specialized adjustments, we constructed an optimized dataset specifically tailored for evaluating large language models (LLMs).

3.3 Dataset construction

Figure 1 represents a JSON data structure, which is an example entry from the dataset. Specifically, it corresponds to an item from the "National Center Test for University Admissions" dataset, indicating its source as a standardized Japanese examination. Key components shown in the image include:

Source: Indicates the origin of the question as the "National Center for University Entrance Examination."



```
{
    "source": "National Center For University Entrance Examination",
    "subject": "Kagaku(main exam)",
    "year": "1993",
    "question_id": "Q14",
    "label": "問２",
    "text": "\n\n次の問い(問１～３)に答えよ。\n\n\n次の記述①～⑤のうちから，正しいものを一つ選べ。\n",
    "choices": {
        "choice1": "①水溶液中での酢酸の電離度は，その濃度が小さくなるにつれて，小さくなる。",
        "choice2": "②純水の電離度は，室温で１×10−7である。",
        "choice3": "③一定温度の酸や塩基のうすい水溶液では，水のイオン積はpＨによらず一定である。",
        "choice4": "④pＨ４の塩酸とpＨ12の水酸化ナトリウム水溶液とを同体積ずつ混合すると，その溶液のpＨは８となる。",
        "choice5": "⑤酢酸水溶液に水酸化ナトリウム水溶液を加えると，溶液中の酢酸イオンの濃度が減少する。"
    },
    "answer_style": "multipleChoice",
    "answer_type": "sentence",
    "knowledge_type": "KS,DM_C",
    "need_image": "yes",
    "image_files": [
        "Center-1993--Main-Kagaku-003.png"
    ],
    "correct_answer": "3",
    "score": "4"
},
```

Figure 1. An example of National Center Test for University Admissions dataset with some shapes and text

Year and Question ID: It states the year as 1993, with a specific question identifier.

Text: Contains the main body of the question in Japanese, related to chemistry concepts involving chemical reactions and ions.

Choices: Lists several options (choices "1" to "5"), each representing possible answers to the multiple-choice question.

Answer Style and Knowledge Type: Defines metadata, indicating the question follows a multiple-choice format, tests conceptual knowledge ("conceptual").

Need Image: When a figure is required for answering the question, the key "need_image" is set to true; otherwise, it is set to false. Additionally, the filenames of corresponding images, if applicable, are included. Supplementary figures or diagrams are not provided for all questions.

Correct Answer and Score: The field "correct_answer" identifies the correct choice ("3"), and "score" assigns a point value ("4").

This structured format suggests that the JSON is intended for automated evaluation systems, possibly in AI-driven educational applications, such as machine learning models designed to solve standardized examinations or to assist in educational content analysis. And this structure facilitates direct integration with LLM training and evaluation pipelines. In this study, we prepared datasets comprising exam questions from several years (as shown in Table 1). Although exams were sometimes administered separately for subjects IA and IB due to curriculum differences, we selected the IB exams for data construction and evaluation because they had higher numbers of test-takers. Supplementary exams were excluded from the scope of this study.

Table 1. Number of questions per year (parentheses represent the number of questions without image)

Year	Physics	Chemistry	Biology
2009	22 (0)	25 (13)	25 (6)
2005 (IB)	21 (0)	29 (8)	27 (0)
2001 (IB)	21 (0)	29 (6)	27 (5)
1997 (IB)	19 (0)	23 (2)	24 (16)
1993	12 (3)	24 (9)	21 (15)

The dataset in this study is available (https://github.com/KyosukeTakami/center-examination-jp).

4. EXPERIMENTS

4.1 LLMs

We evaluate the following three Japanese-oriented large language models (LLMs), each having approximately 13 billion parameters and suitable for inference on a local environment equipped with a single NVIDIA A100 40GB GPU.

- **llm-jp-3-13b-instruct3**
  Developed by the National Institute of Informatics, this model is part of the LLM-jp-3 series, tailored for Japanese language processing [9].
- **google/gemma3-12b-it**
  The Gemma 3 models are trained with distillation and achieve superior performance to Gemma 2 for both pre-trained and instruction fine-tuned versions [3].
- **DeepSeekR1-Distill-Qwen-14B**
  A distilled version of the DeepSeek R1 model, optimized for efficient inference while maintaining strong performance in language understanding tasks [1].

4.2 Text-only LLM

In this study, we evaluated the responses generated by the LLMs using text-only inputs. Although Gemma is capable of multimodal processing, we deliberately provided only text-based inputs—even for questions that ordinarily rely on visual content—to assess its capacity to respond effectively without any image data.

4.3 Prompt

Prompts are translated as follows: for multiple-choice questions,

"質問と回答の選択肢を入力として受け取り、選択肢から回答を選択してください。" (Given a question and multiple-choice answer options as input, please select the correct answer from among the provided options.)

"なお、回答は選択肢の番号（例：1)でするものとします。"

(Note that answers should be provided using the option number (e.g., 1).)

"回答が複数になる場合は（例：3|4)として下さい。"

(If multiple answers apply, please indicate them using the pipe symbol (e.g., 3|4).)

"回答となる数値以外は返さないでください。"

(Return only the numerical value(s) corresponding to your answer. Do not include any additional text.)

We included the following example as a one-shot prompt for solving multiple-choice questions:

{"input": "質問: 日本の首都はどこ？\n 選択肢: 1.東京, 2.大阪, 3.名古屋" (Question: What is the capital of Japan?\n Options: 1. Tokyo, 2. Osaka, 3. Nagoya),

"output": "1"}

4.4 Evaluation

For evaluating LLM performance, this study used the exact match ratio, which measures the proportion of questions where the LLM's predicted answer exactly matches the correct label. This ratio is a widely used evaluation metric for multiple-choice question answering tasks, but it should be noted that it differs from the publicly reported average student scores, where each question has an assigned score and results are normalized to a maximum of 100 points.

5. PRELIMINARY RSULT

5.1 Overall results

Table 2. Overall results

Subject	llm-jp-3-13b-instruct3	gemma-3-12b-it	DeepSeek-R1-Distill-Qwen-14B	Student average score (weighted/normalized score)
Physics 2009	4.55 (--)	**27.27** (--)	13.64 (--)	63.55
2005 (IB)	19.05 (--)	23.81 (--)	14.29 (--)	59.97
2001 (IB)	19.04 (--)	14.29 (--)	9.52 (--)	72.81
1997 (IB)	9.50 (--)	21.05 (--)	**31.58** (--)	70.71
1993	**41.67** (33.33)	16.67 (0.00)	25.00 (33.33)	53.84
Chemistry 2009	17.86 (18.75)	39.29 (50.00)	32.14 (43.75)	69.54
2005(IB)	31.03 (25.00)	37.93 (50.00)	34.48 (62.50)	66.06
2001(IB)	24.10 (16.67)	31.03 (50.00)	24.14 (33.33)	58.51
1997(IB)	17.30 (50.00)	39.13 (50.00)	**34.78** (**100.0**)	62.93
1993	**37.50** (**55.56**)	**45.83** (**66.67**)	29.17 (44.44)	58.69
Biology 2009	24.00 (00.00)	32.00 (**50.00**)	32.00 (16.17)	55.85
2005 (IB)	22.22 (--)	**51.85** (--)	**40.74** (--)	51.58
2001(IB)	**33.33** (0.00)	29.17 (31.25)	**40.74** (45.45)	67.12
1997(IB)	4.17 (0.00)	29.17 (31.25)	37.50 (37.50)	51.73
1993	14.29 (**20.00**)	23.81 (26.67)	38.10 (**53.33**)	59.94

Table 2 demonstrates the results of the evaluation of Local LLMs on our National Center Test for University Admissions Dataset. We provide the exact match ratio for each local LLM. The values in the upper row represent the overall exact match accuracy, including both questions requiring images and those solvable with text alone. Numbers enclosed in parentheses indicate the exact match ratio for questions without accompanying images (i.e., those solvable using text alone). For each subject, the exact match ratio of the best-performing model is indicated in **bold**. Looking at the results for physics (upper rows of the table), all models exhibited exact match ratio below 50%, indicating inferior performance compared to average student accuracy. This is likely because physics questions predominantly require interpreting diagrams (images). As shown in Table 1, only three questions from 1993 could be solved using text alone. In the chemistry results (middle rows of the table), the exact match ratio for text-only questions exceeded 50% for the year 1993 in the llm-jp and gemma3 model, indicating performance comparable to or even surpassing that of human examinees. In the biology results (lower rows of the table), Gemma3 exhibited the highest exact match accuracy among all models and subjects for the year 2005. This performance was considered comparable to human average accuracy levels.

6. LIMITATIONS

One limitation of this study is that questions presumed to require diagrams were evaluated using only textual information. This was done to assess how effectively the questions could be answered based solely on text. In future research, it will be necessary to conduct performance comparisons using multimodal LLMs that incorporate images. For this reason, the dataset prepared in this study also includes image files to support further investigation (see, our dataset: https://github.com/KyosukeTakami/center-examination-jp). Another limitation is that, while we adopted exact match accuracy—a commonly used evaluation metric for LLM evaluation—employing a scoring approach based on question-specific point values might be more appropriate for comparing model performance with human performance.

7. DISCUSSION

Local Large Language Models (LLMs) with approximately 13 billion parameters have demonstrated performance comparable to humans in certain tasks. However, their overall performance often falls short of human levels. To achieve human-equivalent performance across a broader range of tasks, larger models with around 100 billion parameters may be necessary. Advancements like the Petals platform have made it feasible to run such large-scale LLMs on personal GPUs, potentially enabling their integration into educational settings. Since the underlying mechanisms of LLMs are still not fully understood, clarifying what linguistic resources they were trained on and revealing the detailed nature of their errors could help identify more efficient human learning strategies. Moreover, comparing these aspects with human errors may open the door to proposing novel learning methods for humans inspired by the learning processes of LLMs. By comparing the types of errors made by these models with those of human learners, educators can gain insights into learning processes and the underlying mechanisms of LLMs. This approach may contribute to a deeper understanding of both human learning and AI behavior.

8. CONCLUSION

In conclusion, this research provides valuable initial insights into the potential of local LLMs for educational evaluation, demonstrating both their capabilities and limitations. Further work, particularly involving multimodal analysis and a scoring methodology that aligns more closely with human performance metrics, will be crucial in fully assessing the practical applicability of local language models in educational assessments.

9. ACKNOWLEDGMENTS

We thank Fine-tuning and Evaluation WG members of the LLM-jp [8] for their valuable feedback and advice. This study was partly supported by JSPS JP23K17012 and 23H01001.

10. REFERENCES

[1] DeepSeek DeepSeekR1-Distill-Qwen-14B. *https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B*.

[2] Dong, Z., Chen, J. and Wu, F. 2025. Knowledge is power: Harnessing large language models for enhanced Cognitive Diagnosis. *arXiv [cs.AI]*.

[3] Google gemma3-12b-it. https://huggingface.co/google/gemma-3-12b-it.

[4] Han et al., N. 2024. llm-jp-eval: An automatic evaluation tool for Japanese large language models (in Japanese). *Proc. 30th Annu. Meeting of the Assoc. Natural Lang. Process.* (2024), 2085–2089.

[5] Jiang, J., Huang, J. and Aizawa, A. 2024. JMedBench: A benchmark for evaluating Japanese biomedical large language models. *arXiv [cs.CL]*.

[6] Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y. and Radev, D. 2023. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *arXiv [cs.CL]*.

[7] Kurihara, K., Kawahara, D. and Shibata, T. 2022. JGLUE: Japanese General Language Understanding Evaluation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille, France, Jun. 2022), 2957–2966.

[8] LLM-jp et al. 2024. LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. *arXiv [cs.CL]*.

[9] llm-jp llm-jp-3-13b-instruct. https://huggingface.co/llm-jp/llm-jp-3-13b-instruct.

[10] Nejumi leaderboard: https://wandb.ai/wandb-japan/llm-leaderboard3/reports/Nejumi-LLM-3--Vmlldzo3OTg2NjM2.

[11] Todai Robot Project dataset: *https://21robot.org/dataset.html*. Accessed: 2025-03-17.

[12] Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S. and Wen, Q. 2024. Large Language Models for education: A survey and outlook. *arXiv [cs.CL]*.

[13] Wang, W., Chen, W., Luo, Y., Long, Y., Lin, Z., Zhang, L., Lin, B., Cai, D. and He, X. 2024. Model compression and efficient inference for large language models: A survey. *arXiv [cs.CL]*.

[14] Xu, X., Chen, Y. and Miao, J. 2024. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *Journal of educational evaluation for health professions*. 21, (Mar. 2024), 6.